How do I know what my theory predicts?

Article (Accepted Version)

How do I know what my theory predicts?

Zoltan Dienes

University of Sussex

Correspondence:

Zoltan Dienes

School of Psychology

University of Sussex

Brighton

BN1 9QH

dienes@sussex.ac.uk

Abstract

To get evidence for or against a theory relative to the null hypothesis, one needs to know what the theory predicts. The amount of evidence can then be quantified by a Bayes factor. Specifying what one's theory predicts may not come naturally, but I show some ways of thinking about the problem, some simple heuristics that are often useful when one has little relevant prior information. These heuristics include the room-to-move heuristic (for comparing mean differences), the ratio-of-scales heuristic (for regression slopes), the ratio-of-means heuristic (for regression slopes), the basic effect heuristic (for ANOVA effects), and the total effect heuristic (for mediation analysis).

Introduction

Researchers are often interested in the existential question of whether something exists: Should an effect be in a model? Is there an interaction? Are there side effects of the drug? One has to assume something exists in order to estimate it (and in not estimating other things, one presumes that they do not exist). So it would be nice to have a measure of evidence for something existing versus not existing. Significance testing is a commonly used tool for this purpose; however, non-significance is not itself evidence for something not existing. On the other hand, a Bayes factor can provide a measure of evidence for a model of something existing versus a model of it not existing (Etz, & Vandekerckhove, 2018; Morey, Romeijn, & Rouder, 2016). Thus, evidence for existence versus non-existence is put on a symmetric footing. This paper will give practical guidance on using Bayes factors; readers who have no background in their use might want to first read Dienes (2014) and/or Dienes and McLatchie (2018) as introductions congruent with the approach taken in this paper. After introducing the problem of using Bayes factors when there is limited relevant prior information, this paper will provide a number of heuristics for this situation.

A model, as the term is used here, is a representation of the predictions of a theory. The model indicates how plausible different population values are of the parameter postulated to exist (a probability distribution of parameter values being the *model of H1*). The contrast model can simply state the parameter does not exist (this is the *model of H0*). These two models can then be used to calculate a Bayes factor, and hence the evidence for one model versus the other, and thus in this case, the evidence that something exists versus does not. The problem then arises of specifying the sort of effect size the theory predicts, in order to construct a model of H1. This is what many researchers might find difficult. But there can only be evidence that something does not exist given a claim of how big it could be, if it did exist. But how do I know what effect size my theory predicts?

Data collected to test a theory gives information about the size of effect that exists, should it exist. Thus, one might be tempted to use the data for testing the theory to also specify the effect size predicted. But this is double counting, and forbidden by the maths of how a Bayes factor is derived (Lindley, 1991). Or to put another way, in order that theory and data can clash, the model of H1 should not be constructed on the same information that it is tested against. If the same data were used to generate the predictions of a theory as test them, the theory could not be severely tested (cf. Popper, 1963). How can one determine the range of effect sizes consistent with a theory? In order to motivate the problem, first I give an example where there is relevant past research. Then I indicate what types of model will be used for the examples in order to calculate a Bayes factor. Finally, I consider a case where there is no relevant past research, and use that example to show a general approach: The use of heuristics to constrain predicted effect sizes even in the absence of past studies. The bulk of the paper will then describe the heuristics.

Relevant past research helps define the effects expected

Previous work has shown that a two-week mindfulness of breathing intervention increased mindfulness by 0.2 Likert rating points on a mindfulness questionnaire, the FFMQ (Cavanagh, Strauss, Cicconi, Griffiths, et al., 2013). A researcher decides that it would be useful to try as a conceptual replication a two-week mindfulness of walking intervention, given the theory that both mindfulness of breathing and of walking engage the same process, namely mindfulness. She finds a mean difference in her sample of 0.1 Likert units of FFMQ. If she uses the 0.1 mean difference obtained from the sample as the basis for constructing her model of H1, she has double counted the obtained mean difference: Once for forming the predictions, and twice for testing them. A pseudo-Bayes factor constructed from a model of H1 in this way, whose predicted mean effect is chosen to be the same as the data, puts a theory at least risk of being shown wrong.

Instead, the researcher could use the theory that focusing on breath and focusing on walking promote mindfulness in same way (they are both examples of mindfulness training). She could then use the past study on mindfulness of breathing to predict the effect size for the mindfulness of walking study. Note that the theory that two things belong to the same class does work in making that prediction. Hence, the theory can take credit (or blame) in terms of the evidence for this H1 versus H0; the theory can be tested. In general, an important question is when a theory takes credit when a particular model of H1 is tested. A common case is precisely that illustrated here: A theory can take credit (or blame) when the theory claims two things belong to the same class, and that claim is used to construct H1. But often a researcher does not know what prior studies are relevant, or thinks none are. How does one construct a model of H1 then? This is the problem the current paper addresses. Before getting to the main part of the paper we will first consider the sort of models of H1 we will use, and then we consider an example to motivate the main issues to be discussed.

Models

To simplify discussion we will mainly use a very common model of H1 for Bayes factors, which consists of a distribution centred on zero, and the problem is to determine the approximate size of effect predicted, i.e. its scale factor; half of the distribution may also be removed (below 0) to represent a theory making a directional prediction. It may not be apparent why the mode of the distributions is set at zero. This represents in a simple way that smaller effect sizes are more probable than larger ones, which can be useful given a literature that habitually overestimates effect sizes. We will mostly use a half-normal distribution, and the problem is to determine its standard deviation (see e.g. Dickey, 1973; Dienes & McLatchie, 2018). The standard deviation is set to the rough scale of effect expected. Thus, the problem of specifying the model of H1 reduces to specifying the effect size expected. We will notate a Bayes factor based on a half-normal distribution with a mode of 0 and a standard deviation of r, as $B_{HN(0,r)}$. The Dienes (2008) calculator can be used with this model of H1; one needs in addition the effect size and its standard error as a sufficient summary of the data. Another commonly used distribution is the Cauchy (or half-Cauchy) distribution (JASP: van Doorn, van den Bergh, Bohm, Dablander, et al., 2019;

Rouder, Speckman, Sun, Morey, et al., 2009); in the same way, the problem is to set its scale factor, i.e. to determine the rough scale of effect expected. For convenience the term "scale factor" will be used to refer to both the scale factor of a Cauchy and the standard deviation of a normal. We will notate a half-Cauchy with a mode of 0 and a scale factor of r as $B_{HC(0,r)}$. For the same scale factor, the normal and the Cauchy distributions give very similar Bayes factors, with the Cauchy slightly favouring H0 relative to the normal distribution (Dienes, 2017a) (see Box 1). None of these models may be appropriate in any given case (e.g. see Dienes, 2014; Gronau, Ly, & Wagenmakers, in press); however the models are good enough approximations sufficiently often that they will serve as good vehicles for discussing what this paper will focus on (see Dienes & McLatchie and Rouder et al. for justifications of these models). For the sake of discussion, we will treat a $B > 3$ as good enough evidence for H1 over H0, a $B < 1/3$ as good enough evidence for H0 over H1, and a B in between those values as being non-evidential (cf. Jeffreys, 1939). However, there are no real cut offs, these are just rough guidelines adopted because in practice decisions often have to be made. Further, we as a community may (and I think should) decide that 3 and 1/3 are not good enough evidence for many scientific problems: Schönbrodt, Wagenmakers, Zehetleitner, and Perugini, (2017) recommended a cut-off of at least 5 (or 1/5); *Cortex* Registered Reports uses 6 (or 1/6) (Cortex, 2019); and Benjamin, Berger, Johannesson, Nosek et al. (2018) recommend 20 (or 1/20) for one-off studies. If the same community is also using significance testing, the results of the two approaches can be aligned as best as is possible (even though there is no monotonic transform between Bayes factors and p-values) by using a cut off of 3 for Bayes factors for a community also using 5% significance; of 6 for a community using 2% significance; and 20 for 0.5% significance.

Box 1

Normal vs Cauchy distributions for Bayes factors

Sometimes models of H1 employ a Cauchy distribution (e.g. Rouder et al., 2009); sometimes a normal (Dienes & McLatchie, 2018). The function of the model of H1 is to represent the predictions of a theory simply and adequately. What about the normal and Cauchy distributions is important in distinguishing them for constructing models of H1? Consider models where the mode of these distributions is set to be zero. About 5% of the area of a normal (or half-normal) distribution is beyond two standard deviations; so two standard deviations is a rough maximum for a normal distribution. About 5% of the area of a Cauchy (or half-Cauchy) distribution is beyond seven scale factors; so about seven scale factors is a rough maximum for a Cauchy distribution. Turning this around, if one had a reason for setting the rough plausible maximum effect that could be obtained as max, then using a half-normal distribution set SD = max/2 (Dienes, 2014). However, if one used a Cauchy, set the scale factor to max/7. The difference depends on the scientific case for the relation between the expected value and the maximum value. In the absence of any information about this relation, use the half-normal distribution, because this spreads out the uncertainty to represent that lack of information. If one had some information that the effect size would be small relative to the maximum (roughly 1/10 to 1/5), use the half-Cauchy. We will see how this plays out in the examples that follow. Mainly the half-normal will be used, because this does not presume additional information restricting the expected size of the effect.

Example with no relevant past research

Now we will consider an example to motivate the bulk of the paper. Theory A claims that autistics will perform worse on a novel task than controls. Theory B claims that autistics and controls will perform the same. The task has a chance baseline of 0% and a maximum score of 50%. The results are autistics score 8% above baseline (SE = 6%, N = 30) and controls score 10% (SE = 5%, N = 30). The difference (2%) is non-significant, $t(58) = 0.25$, $p = 0.80$, Cohen's $d = 0.05$. One reaction might be that the non-significance of the result means theory B is supported. But a non-significant result does not distinguish evidence for H0 over H1 from not much evidence either way. To know if there is evidence for H0 over H1, we need to know what size effect we could be trying to pick up. In other words, how should we model H1? One temptation might be to use a "default" model of H1, for example the model given by default in JASP for a t-test, a "JZS" Bayes factor, i.e. a Cauchy with a scale factor of 0.7 Cohen d units. This gives $B_{C(0,\ 0.7\ \text{Cohen's d units})} = 0.27$. On the face of it, we have evidence for Theory B, because the Bayes factor is less than a 1/3. But there is no such thing as a

default theory. So there cannot be a default model of H1 (Etz, Haaf, Rouder, & Vandekerckhove, in press; Lee & Vanpaemel, 2018; Rouder, Morey, Verhagen, Province et al., 2016). In fact, the data themselves give information that the conclusion (that there is good evidence for theory B over theory A) was too rash. The control group score 10%, so on theory A, which says autistics will perform somewhere between the control and 0, the difference between the autistics and the control group cannot be more than about 10%. Thus, modelling H1 as a half-normal with an SD = max/2 = 5%, gives $B_{HN(0,5\%)} = 0.94$: No evidence one way or the other. This is clearly a reasonable conclusion because the standard error of the difference is 7.8%, about as big as the maximum possible difference. There cannot be evidence for whether or not there is a difference, if the standard error of the difference is as large as the maximum plausible difference. One can think of this as a floor effect on the difference score: If the sample difference cannot be higher than the standard error of the difference there is a floor effect.

Notice the control group mean was used to inform the maximum difference between the autistics and control group. How does this relate to the principle stated earlier, that the mean difference itself cannot be used to predict the same mean difference? In this case, information other than exactly that tested was used (albeit the information used was correlated with that tested). The information used constrained inference in a plausible way: floor effects appropriately became non-evidential (see Figure 1a). Further, if there had been no floor effects, using the control group to define a maximum difference still gives full scope for either theory to clash with data, to be shown wrong: If there had been a small SE of difference, the autistic group could have been close to control group, Figure 1b, thereby refuting theory A; or else close to chance, Figure 1c, thereby refuting theory B. Thus, we have cheated information out of the data in a useful way that does not prejudice against theory testing. That is, the procedure can provide what Popper (1963) called a severe test of relevant theories. A severe test is one where a theory is made to stick its neck out; if the theory is wrong, it can easily be found wrong.

Box 2 Raw vs standardized effect sizes effect sizes

It may be tempting to believe it is easier to set an expected effect size for a theory using standardized rather than raw effect sizes. Standardized effect sizes, such as Cohen's d, remove the units of measurement (seconds, Likert units, etc) and so render the units irrelevant. It may seem this makes there less to think about, and so the problem is easier. However, standardized effect sizes are signal to noise ratios, and theories and practical claims are usually about signals, and not the noise through which they are measured. A slimming regime is effective if it produces a certain number of kg loss on average, regardless of the random error in the scales. In fact, focusing on standardized effect sizes can lead to misleading conclusions. If one is motivated to conclude an effect does not exist, one could measure it with few trials, as the population standardized effect size over participants is then small (for an example, see Dienes 2017b, 24 – 30 minutes).

One way of seeing why the "default" Bayes factor was not reasonable in the example on autism is that in this case a Cohen's d of 0.7 corresponds to a raw difference of 19%. So the default model of H1 is predicting an effect of around 19% and possibly as large as 19% × 7 = 133% (see box 1). This is clearly unreasonable for the study.

When effect sizes are considered in raw units, they are often easier to evaluate (Baguley, 2009). The greater ease of working with raw rather than standardized units is a point (which many may find counter-intuitive) that this paper will build on. For example, the ratio-of-scales heuristic and ratio-of-means heuristic will illustrate how thinking in terms of raw regression slopes can be easier than in terms of Pearson correlation coefficients. Most generally, if we care about the units in which we measure things (which as scientists we should), a corollary is we should learn to think in those units and not throw them away the first chance we get. As testing existential claims requires a scientific judgment about the sizes of effect that might obtain, such testing is at least a matter of science as statistics.

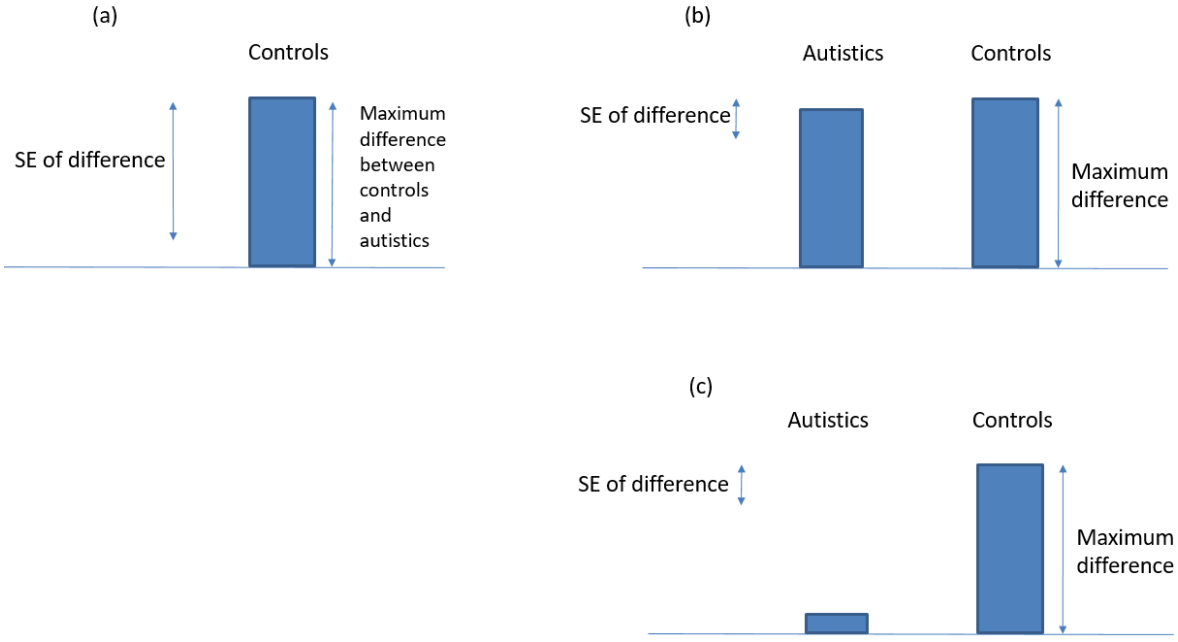

Figure 1.

Using the control group to define the maximum difference between the experimental group and control group can have useful inferential properties. In (a), the procedure can show that there is an effective floor effect, in the sense that the standard error is as large as any difference that could be expected, so no sample difference can be far enough away from the floor defined by the standard error of the difference. Yet, if the SE were small enough, the actual difference between groups could strongly count against either a theory that predicted a difference (b), or a theory that predicted no difference (c): Theory can still clash with data.

Box 3 Robustness checking

For all statistical analyses, including the use of Bayes factors, it is worth considering how robust conclusions are to reasonable changes in assumptions. With Bayes factors testing a theory, the model of H1 represents the predictions of the theory. But there could be several equally good ways of modelling H1 to represent the same theory. Thus, the conclusion would only be robust if most of the models came to the same qualitative conclusion. Dienes (2015; Appendix 12.1) used different distributions (uniform vs normal vs half-normal) to show the precise shape of the distribution can make little difference to conclusions in individual cases when the distributions represent roughly the same scientific assumptions. Given just one form of distribution, such as a half-normal, the conclusion is robust if the range of scale factors (standard deviations) leading to the same qualitative conclusion roughly span or contain the range of scientifically plausible values. JASP produces a graph showing the Bayes factor for different scale factors. When interpreting this graph, the issue is not whether all Bayes factors agree in the conclusion implied, but whether the range of scientifically plausible scale factors is roughly contained in a range of scale factors giving the same conclusion. This notion is formalized in the robustness region, which is a type of mini multiverse (Steegen, Tuerlinckx, Gelman, Vanpaemel, 2016). There are no precise rules yet to say how robust is robust enough; and probably there should not be. But if the robustness region is always provided, any reader can determine if their preferred rough scale factor is contained in the region. If a conclusion is not robust enough, in principle more data can be collected until the conclusions are more robust. With Bayes factors, it is fine to continue collecting data until the evidence is good enough (Rouder, 2014, 2019). The robustness regions in this paper were calculated by iteratively entering different scale factors in the Dienes (2008) Bayes factor calculator until the limits of good enough evidence were obtained (cf Mclatchie, 2018).

One way to ensure some robustness is to use a stopping rule for achieving a degree of evidence that clearly exceeds what is taken to be good enough. For example one may run participants until the Bayes factor is greater than 10 or less than 1/10, and then report the robustness region with respect to cut offs of 5 and 1/5.

This paper will generalize the above considerations and present a set of heuristics for obtaining a ballpark estimate for a reasonable predicted effect size. We will present for each Bayes factor a "Robustness Region" notated as "RR [min, max]" where min is the minimum scale factor that leads to the same qualitative conclusion (i.e. good evidence for H1 over H0, if $B > 3$; good evidence for H0 over H1 if $B < 1/3$; and not much evidence at all otherwise);

and max is the maximum scale factor that leads to the same conclusion[1] (see Box 3). (Note: If the conclusion is not much evidence at all, min will always be 0; and if the conclusion is good evidence for H0 over H1, max will always be infinity, ∞. If a scale has a maximum, e.g. a 0-7 Likert scale so the maximum difference is 7, if the maximum exceeded 7, one could notate as ">7".) None of the heuristics are guaranteed to produce sensible answers in context; scientific judgement is always needed for all aspects of model building, including this one. Nonetheless, the heuristic can do its job merely if it puts one in the right ball park; if the robustness region is about the width of a ball park then the conclusion is safe. (That is, if the range of scientifically plausible scale factors is contained with the robustness region, the conclusion is safe.) The heuristic will often also help us see what range of scale factors is scientifically plausible, as we will see below. In the example we have been using, the robustness region for a half-normal distribution model of H1 is $RR_{1/3>B>3}$ [0, 28%]. That spans the whole ball park (a standard deviation of 28% corresponds to a plausible maximum difference of 56%), so the conclusion of insensitive evidence is safe. To remind the reader, the scale factor is the key aspect of the model indicating roughly how big the population difference is between autistics and controls; a given scale factor indicates that a plausible population difference lies between 0 and about twice that scale factor (for a normal or half-normal distribution).

Now we consider the heuristics. First the *room-to-move heuristic*. Second, the *ratio-of-scales heuristic*. Third, the *ratio-of-means heuristic*. Fourth, the *basic effect heuristic*. Finally, the *total effect heuristic* for mediation.

The room-to-move heuristic

The example used previously with the autistics versus the controls motivates the heuristic of using one condition to define the rough maximum difference that could be obtained between conditions: The one condition informs us of how much room there is left to move for the other condition in order to satisfy the constraints of the theory. We now consider a real example. Consider the theory that people pursue relationships for a mix of eroticism and nurturance. In a polyamorous relation one can have different partners for different needs; thus, each partner might satisfy the particular need they are assigned to better than in a monogamous relationship, where one partner has to satisfy all needs. Balzarini, Dharma, Muise, and Kohut (2019) investigated the quality of polyamorous versus monogamous relationships. On a 1-7 scale of how nurturing their primary partner was, people in monogamous relationships rated their partner's nurturance as 5.85. For polyamorous people I first will pick a subgroup to make the example interesting. Taking polyamorous people without a self-defined primary partner, when relationship length was controlled for, the mean nurturance rating for the partner they mainly lived with was 5.80, SE of difference was 0.11, $t(\approx 2500)$ = 0.42 (see Table 5 of Balzarini et al). This is non-significant. But non-significance does not mean there is evidence for no difference. To define the evidence, the scale of effect predicted by H1 needs to be determined. How should H1 be modelled? Given the monogamous group score, how different could the

[1] Thanks to Balazs Aczel for suggesting this concept of a robustness region.

polyamorous group be? Given the monogamous group scored 5.85, and the top of the scale is 7 units, the maximum increase is about 1.15 units (see Figure 2), that's the room to move for the polyamorous group. So to model H1 using the room-to-move heuristic, use a half-normal distribution with SD = max/2 = 0.58 rating units. This gives $B_{HN(0,.58)} = 0.13$, $RR_{B < 1/3}$ [0.22, >6], evidence for H0 over H1. We can assess the robustness of the conclusion by taking into account information from the other polyamorous couples, i.e. with defined primary and secondary partners. In this case, the polyamorous couples rated their partner they mostly lived with as being more nurturing than monogamous couples did by 0.57 units (SE = 0.10). 0.57 is thus a more informed estimate of the sort of difference that could be expected. 0.57 is very similar to the result given by the room-to-move heuristic (a similarity that cannot in general be guaranteed) and, importantly, well within the robustness region.

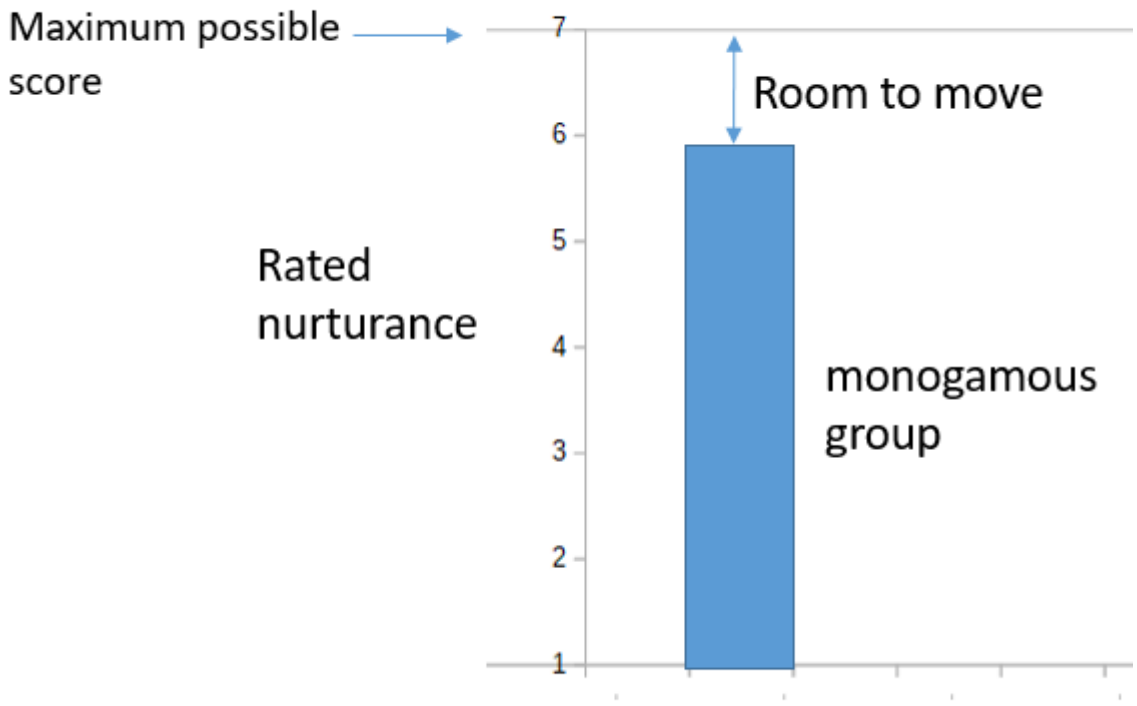

Figure 2

The polyamorous group has 1.15 units room to move in partner's nurturance, given the monogamous group's mean, and given that 7 is top of the scale. Thus, one group can provide constraints on what difference can be found between groups, defining the room-to-move heuristic.

One might ask: Why not take the polyamorous group's mean as given and see how much room there was for the monogamous group to be lower than that? In that case, the room to move would be about 4.85 (from 5.85 to the bottom of the scale, 1), and that could in principle make a difference to conclusions (though in fact not in this case). Choose the way

round that gives the smallest room to move, because that will show up any floor or ceiling effects. So in this case, it was justified to go with the monogamous group as setting the room to move.

The room-to-move heuristic is based on a point estimate from one group. For example, we have assumed that the estimate of nurturance of 5.85 for the monogamous group is a reasonably precise estimate. This has the advantage of simplicity. But it also disregards the uncertainty in that estimate. In this case, the standard error of the estimate is 0.03 nurturance units, so the estimate is precise enough. The function of the heuristic is to put us in the right ball park; the question then is the width of the robustness region. The robustness region in this case includes all reasonable rooms-to-move.

In order to analyse interactions, Gallistel (2009) suggested taking a key simple effect as the maximum size the difference in simple effects could be, i.e. as the maximum size of the interaction (a strategy later recommended by Dienes, 2014). For example, highly hypnotisable people are tested on the Stroop test either with no suggestion, or a word blindness suggestion that the words on the screen are written in a meaningless foreign script (Raz, Shapiro, Fan, & Posner, 2002). The suggestion has been found to reduce the Stroop interference effect. Consider the first time the study is run so there is no prior information about how effective the suggestion should be. In the no suggestion condition Raz et al. found that RTs for incongruent and neutral words were 860 ms and 748 ms respectively. In the suggestion condition they were 669 and 671 ms respectively. In the no suggestion condition the interference effect therefore was (860 ms - 748 ms) = 112 ms. That is the simple effect of word type for no suggestion. In the suggestion condition the interference effect was (669 – 671) = -2 ms. That is the simple effect of word type for the suggestion condition. Given that the interference effect with no suggestion is 112 ms, the most suggestion could plausibly reduce interference is therefore about 112 ms. That is the only room in which it has to move. Therefore we could model the H1 for the interaction word type X suggestion as a half-normal distribution (directional: suggestion should reduce not increase the interference effect) with a standard deviation of max/2 = 112/2 = 56 ms. So we have predicted the size of effect. In fact, the raw interaction effect was 112 – (-2) ms = 114 ms. Now to find the standard error of the effect: The interaction test reported in the paper was $F(1, 30) = 29.35$, which corresponds to $t(30) = \sqrt{29.35} = 5.42$. Therefore, the SE for the interaction = (raw effect size)/(obtained t) = 114 ms/5.42 = 21 ms. The Dienes (2008) calculator can be used with this information, $B_{HN(0,56)} = 2.86 \times 10^5$, $RR_{B>3}$ [4.3, 4×10^4], evidence for suggestion reducing Stroop interference, with the robustness region containing all remotely plausible scale factors. (In fact, a meta-analysis by Parris, Dienes & Hodgson, 2013, indicated that the suggestion roughly halves the interference effect, so the model of H1 for the interaction we now pre-register based on past data is precisely also that which would be given by the room-to-move heuristic, Palfi, Parris, McLatchie, Kekecs, et al., 2018). Note here the advantage of using raw units, milliseconds. If fewer trials of the Stroop test were run, the expected standardized effect size would change. But the fact that the effect of suggestion is approximately to halve the raw interference effect would remain invariant.

Ratio-of-scales heuristic

The ratio-of-scales heuristic may be useful when correlating or regressing one variable on another. The task is to determine if a simple version of the theory tested could make a correspondence between two low points on the scales and two high points. Notice the task is not to determine the spread of the data for each variable, but to determine what a simple theory would predict given the meaning of the scale points. The principle can often be applied to regressions. Lush et al (2019) asked people to estimate the time that a tone occurred. The tone happened 250 ms after a button press. According to a theory we developed (Bayesian cue combination theory applied to time estimation), the experienced time of the tone should be pulled towards that of the button press according to a "relative precision" which varied from 0% to 100%. One way to test the theory would be to determine if the shift in the estimated time of the tone correlated with the relative precision: The theory predicts that the higher precision, the greater the shift. What size correlation could we expect? r= 0.2? 0.6? 0.8? Who knows. If we think in terms of raw units, prediction becomes easier. The maximum the shift in timing could be is if the tone was shifted all the way over to the button press, i.e. a shift of 250 ms. In the simplest version of the theory this is what would happen with a relative precision of 100%. According to the theory, there would be no shift with a relative precision of 0%. So the raw slope of shift against precision in this case is the length of the scale for shift (250 – 0 ms) divided by the length of scale for precision (100 – 0%), i.e. 2.5 ms per percent, the ratio of the scales (see Figure 3). This is the maximum slope that would occur if the only mechanism was the one postulated and it operated with complete effectiveness. Thus, the ratio-of-scales heuristic gives the maximum slope that could be expected. Hence we can model H1 of the raw regression slope with a half-normal distribution with SD = 2.5/2 = 1.25 ms per percent[2]. In fact, Lush et al obtained a raw regression slope b = .59ms (SE= .26), $t(68) = 2.23$, p= .029, $B_{HN(0,1.25)} = 4.74$, $RR_{B>3}$ [0.16, 2.1]. The robustness region goes from very small to almost the maximum slope plausible, so the evidence for there being a slope is robust to the value of the scale factor in this case.

[2] In fact, a more realistic theory predicts a much smaller shift even for maximum precision. One could deal with this simply by using a half-Cauchy distribution with scale factor 2.5/7 = 0.36 ms per percent. As it turns out either scale factor is in the same robustness region, so it makes no difference.

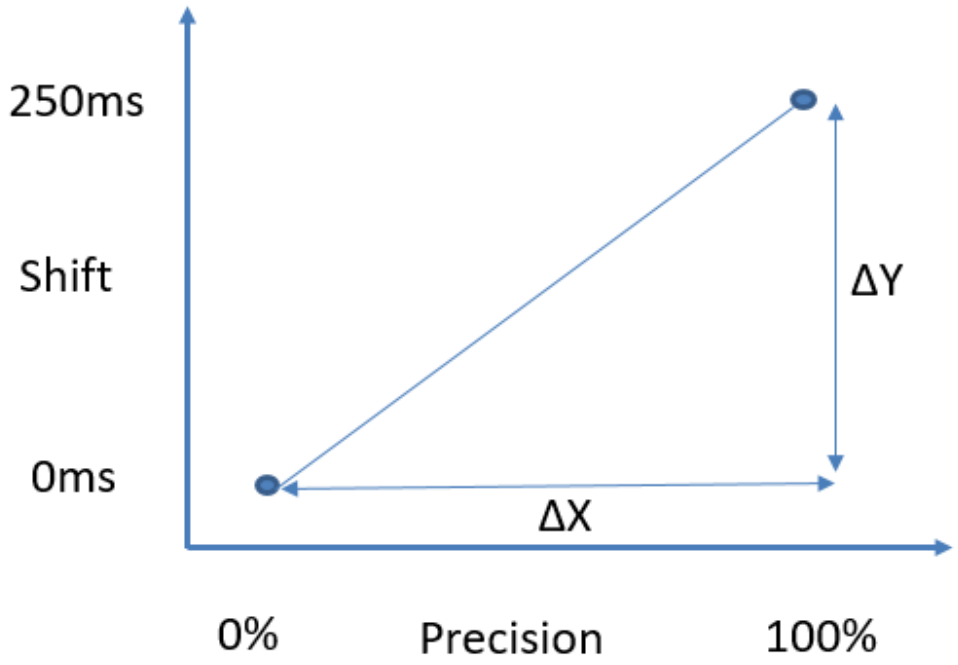

Figure 3

Illustration of the ratio-of-scales heuristic for regression (or correlation). The maximum slope the theory in Lush et al (2019) predicts is the ratio of the lengths of the two scales, i.e. 250 ms / 100% = 2.5 ms per percent.

Monin et al. (2017), using construal theory, predicted that "that women who were high in marital satisfaction would experience the greatest distress on days when they perceived more than average levels of partner suffering. This is compared with men and women low in marital satisfaction and men high in marital satisfaction." The DV was marital distress measured on a 1-4 scale ('not at all stressed' to 'very stressed'). The IV was perceived physical suffering of partner on a 1 ('did not suffer') to 10 ('suffered terribly') scale. If distress increased with suffering in a simple way, and subjects used most of the scale points a fair amount of the time, 'no suffering' would be the response to 'no distress', i.e. the relationship would start with (1,1) and finish with 'terrible suffering' going with the most distress, i.e. the line would go through (10,4). To a first approximation, the slope would be expected to be (4-1)/(10-1) = 0.33 distress units per suffering units. But any variable that affected distress independently of suffering would reduce the relationship. The *ratio of scales heuristic* is to treat the ratio of the scale ranges as a rough maximum. That is we would model the H1 for the relation of distress to suffering as a half-normal with an SD of 0.17 (half the maximum). The authors believe that the relationship between marital distress and partner suffering will hold well for partners with high marital satisfaction but not for those with low satisfaction[3]. For high satisfaction males, *b* = .03 distress units per suffering

[3] They obtain a significant distress vs suffering slope for high satisfaction females, estimated from the graph as 0.22 distress rating units per suffering rating unit. (Notice the ratio of scales heuristic provides a scale – 0.17 distress units per suffering unit - in the right ball park in this example.) As the authors do not give an exact *p*,

unit (estimated from graph), SE = .02. $B_{HN(0,0.17)}$ = 0.66, non-evidential, $RR_{1/3>B>3}$ [0, 0.35]. The maximum scale factor in the robustness region is high given our arguments that a plausible maximum is around 0.33; so the conclusion is robust. Therefore the conclusion in the abstract that "men who were high in marital satisfaction experienced heightened daily distress irrespective of their perceptions of level of spousal suffering," is not supported if "irrespective" is read as meaning 'no relation with suffering'.

Ratio-of-means heuristic

Some scales, for example reaction times or d' (discrimination), have no obvious high point on the scale to relate to a high point of another variable. It may then be difficult theoretically to a priori fix a plausible correspondence between the two scales. Salvador, Berkovitchd, Vinckiera, Cohen, et al. (2018) regressed a measure of thought suppression (in units percentage correct) against ability to discriminate whether a no- think cue was present (in units of d'), with the latter measure taken to be a measure of conscious perception. The raw slope was -5.7 percent per d' unit[4] $t(42) = 0.77$, $p = .45$, "indicating that people's ability to discriminate masked cues did not predict their memory effect" (pp 194-195), thus indicating the thought suppression (memory effect) was triggered unconsciously. The non-significant result does not justify the conclusion of no relation between thought suppression and conscious perception. (There are arguments against first order d' being a valid measure of conscious perception, Dienes & Seth, 2018; but the authors assumptions can be accepted for the sake of determining what tests would be relevant for those assumptions.) What strength of relation could be predicted if both measures depended on conscious perception of the cue? d' goes from 0 to infinity. What high level of d' should correspond to a high degree of thought suppression? The ratio-of-scales heuristic is hard to apply in this case. But we may use a ratio-of-means heuristic. The *ratio-of-means heuristic* is akin to the room-to-move heuristic being applied to each variable. Given that the mean suppression was 6% and the mean d' was 0.35, on the theory that both depend on a single knowledge base (e.g. conscious perception), then they should go to zero together (see Figure 4). Thus, on the theory, the slope should be the ratio of the means, 6%/0.35 = 17 percent per d' unit. This is a maximum because it assumes that all systematic variance is due to conscious knowledge. Thus, we can model H1 as a half-normal distribution with SD = 17/2 = 8.5% per d' unit. With these assumptions, $B_{H(0, 8.5)}$ = 0.43, $RR_{1/3<B<3}$ [0, 12], indicating the data are non-evidential. The robustness region reaches a moderately high value of the slope (given an estimated maximum of 17), indicating that the conclusion (that there is not enough evidence) is

we cannot get an exact SE, but there is no doubt a Bayes factor would give good evidence. For the sake of argument take p < .001 as p = .001; this gives $t(40) = 3.06$ (*df* = 40 is a very rough guess based on the smallest df in their table – but the issue is what could be done in principle, and the answer will be roughly right). So SE = parameter/*t* = 0.22/3.06 = .07 distress units per suffering units. $B_{HN(0, 0.17)}$ = 51.86, $RR_{B>3}$ [0.027, 6.50]. Given the arguments for a plausible maximum of 0.33, and no grounds for thinking the effect below .05, the conclusion is robust that there is evidence for H1.

[4] The authors report mean suppression as 6%, and mean d' as 0.35, and the intercept as 8%. Thus, the slope is (6- 8)%/0.35 = -5.7 % per d' unit. It has a standard error of slope/*t* = 5.7/0.77 = 7.4 percent per d' unit.

somewhat robust to scale factor[5].

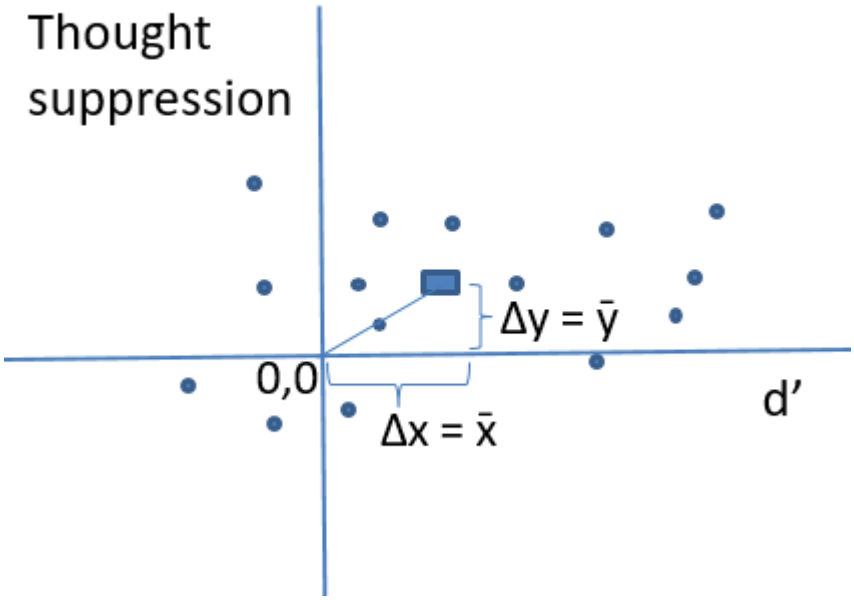

Figure 4

The ratio of means heuristic. For these imaginary data points, let the rectangle mark the mean level of thought suppression and mean level of d'. Based on the theory that both variables depend on a single knowledge base and should thus go to zero together, the expected slope is the ratio of the means. The Y-axis is a difference in percent correct between two conditions so has a true zero; d' has a true zero, namely when discrimination is at chance.

The basic effect heuristic

One can often take the size of a basic effect as a rough scale for how much that effect could be manipulated. Martin and Dienes (in press) used this principle to test whether different types of hypnotic induction were differentially effective in changing response to suggestion. If people were given 10 hypnotic suggestions, and coded as having passed or failed the suggestion (i.e. whether or not they sufficiently experienced the suggested effects), what effect do different inductions have on pass rate (number of suggestions passed out of 10)? The scale factor for the model of H1 for the difference between different inductions was set as the difference between no induction and the standard induction. That is, the bigger an effect any induction has on response, the more inductions may differ between themselves, in the same way as adult shoe sizes differ more between themselves than baby shoe sizes. If people are given 10 hypnotic suggestions, each of which could be passed or failed, a standard hypnotic induction versus no induction increases pass rate by 1.46 suggestions. Thus, the scale factor for the difference between different inductions was set at 1.46

[5] A problem with this regression is the error in measurement of d'. Simone Malejka is working with me (and Miguel Vadillo and David Shanks) to come up with a simple Bayesian solution to this problem (cf Matzke, Ly, Selker, Weeda, et al., 2017).

suggestions out of 10. An "indirect" induction had been argued to be especially powerful; past research showed a difference between standard and indirect inductions of 0.01 passes (SE = .25). This gives $B_{H(0, 1.46)} = 0.20$, $RR_{B<1/3}$ [0.9, >10], evidence that the indirect induction is no different than a standard induction on average. Ziori and Dienes (2015) investigated how gender and attractiveness of facial stimuli may affect implicit learning of sequences of those stimuli. The average level of implicit learning above baseline (6%) was taken as a rough scale by which that effect could be modulated by the manipulations, and used as the scaling factor for all effects in the three-way 2X2X2 ANOVA used (every 1-degree of freedom effect, whether main effect, interaction or simple effect, can be expressed as a contrast in raw units). Casper, Desantis, Dienes, Cleeremans et al. (2016) used the height of an ERP component as the maximum that the component could be modulated (based on past experience with how much such components are typically modulated).

One could broaden the heuristic further to a *reference effect heuristic*, whereby the size of one effect on a study is used as a basis for expecting the size of another effect (perhaps multiplied by a constant, cf.Palfi et al., 2018, discussed above). For example, in fMRI, one could use a standard contrast to define the effect expected for a contrast of interest. The amount of conscious perception in a conscious condition provides the expected effect in the (potentially) subliminal condition, if it were actually based on conscious perception (Dienes, 2015). If a previous experiment used RTs and the current study is using d', there may be a standard effect you could use to convert RTs to d' (cf. Dienes, 2014, supplemental materials, section 2).

The total effect heuristic for mediation

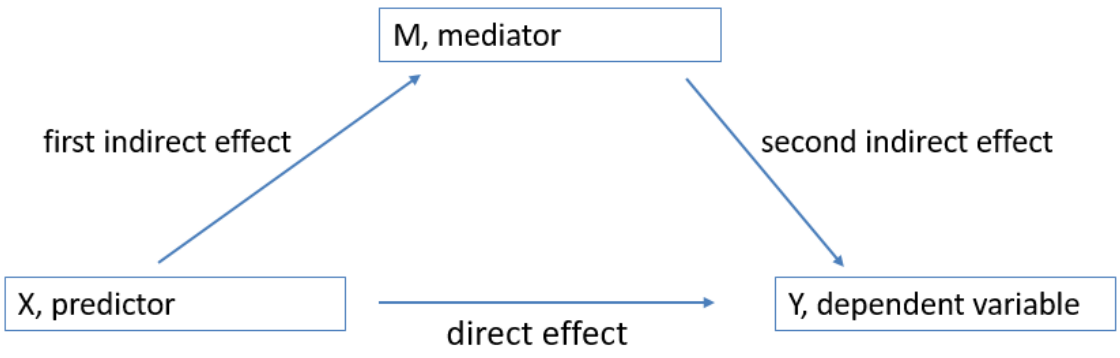

Figure 5

A model of X causing Y to some degree through M, the mediator. The equations defining the paths are: $M = a_1 + b_1 \times X$; $Y = a_2 + b_2 \times X + b_3 \times M$; and $Y = a_3 + b_4 \times X$. The a_i terms indícate that the regresión slopes are in raw units. Given these equations the meaning of the effects are: b_1 = first indirect effect, b_3 = second indirect effect, indirect effect = $b_1 \times b_3$; b_4 = total effect, b_2 = direct effect

In a mediation analysis one might want to know whether the effect of X on Y is mediated completely, partially, or not at all by M (see Figure 5). By frequentist methods, evidence for some mediation can be provided by the first and second indirect effects each being significant (the method of joint significance, e.g. Woody, 2011). Recently, Yzerbyt, Muller, Batailler, and Judd (2018) argued that this method should be preferred to the currently more common use of a single index of the indirect effect. Whatever the method, the main problem for frequentist methods comes when trying to get evidence for full mediation or no mediation, because each of those claims depends on evidence for an H0. We can rephrase the method of joint significance in terms of Bayes factors (cf. Nuijten, Wetzels, Matzke, Dolan et al., 2015, for a different approach). Thus we have for the indirect effects: (i) if $B < 1/3$ for either indirect effect, then there is evidence for no mediation; (ii) if $B > 3$ for both then there is at least partial mediation; and (iii) if either are insensitive AND the other $B > 1/3$, then there is no evidence either way. These tests are regressions and therefore the ratio-of-scales or ratio-of-means heuristics may provide models of H1.

Now take the case of testing for full mediation. Assume there is evidence for an indirect effect. Then, for the direct effect (i) if $B > 3$ then there is not full mediation; (ii) if $B < 1/3$, then there is full mediation; and (iii) if $1/3 > B > 3$, then there is no evidence either way about full mediation. In testing the direct effect, there is a simple heuristic that can be used. Mathematically, total effect = direct effect + indirect effect. Thus, one may have a theory in which the total effect is the maximum that could be expected for the direct effect[6]. To test this theory, model H1 for the direct effect using the uniform [0, total effect]. This is the total effect heuristic. (We use a uniform in this case because there is typically no reason to expect that the direct effect will be closer to either 0 or the total effect.)

Consider a study in which openness to experience (X) is used to predict relationship satisfaction (Y), with the mediator richness of fantasies (M), all rated on Likert scales. The total effect is 0.10 Likert unit of Y per Likert unit of X (SE = 0.02), $t(450) = 5.00$, $p < .001$ (that is, X predicts Y), and the direct effect is 0.04 (SE = 0.03), $t(450) = 1.33$, $p = .18$ (that is, how well X predicts Y when M is partialed out). A typical but incorrect temptation may be to conclude from a significant total effect and a non-significant direct effect that there is complete mediation: openness to experience only increases relationship satisfaction via increasing the richness of fantasies. Indeed, the direct effect is not just non-significant, the JZS default Bayes factor for the direct effect is $B_{C(0,r=0.35)} = 0.11$, i.e. evidence for H0, seeming to confirm the claim of complete mediation. But the maximum that the direct

[6] This is a theory and not a mathematical inevitability because the indirect effect may be negative (cf. Pearl, Glymour, & Jewell, 2016).

effect could be (on the theory that openness increases fantasy richness which increases relationship satisfaction) is the total effect, i.e. .10 Likert units per Likert unit. Using the total effect heuristic, for the direct effect $B_{U[0,0.10]} = 1.62$, $RR_{1/3<B<3}$ [0, 0.5][7]. Thus, the data are non-evidential, and robustly so over any plausible upper limit for the uniform.

Discussion

A scientist tries to explain the world. The explanations can be tested via their predictions. For this, we need a model of the predictions, minimally the sort of effect size, ideally in raw units, that is expected to occur. Even without prior work in the field, there are heuristics that enable minimal constraints on what can be expected. So long as these constraints put one in the right ball park, and help define what the ball park is, evidential conclusions follow if they are robust to about the width of a ballpark. Notice that the Bayes factors this paper have used do not involve H1s with point predictions; they respect the vagueness of real psychological theory in representing a range of possible effect sizes. In considering robustness, we are making sure that the plausible range of the width of that plausible range leads to similar conclusions.

There are no strict default effect sizes in theory testing, hence no objective or default Bayes factors (see Box 4). A proposed default Bayes factor is not an invitation to stop thinking; it is an invitation to think whether the suggested scale is relevant to the problem in context. In many cases suggested default values (e.g. Cohen's d = 0.7) may fall in the same robustness region as a Bayes factor informed by scientific context. But there is only one way to find out; one has to consider what scientific constraints there are and see what they imply.

This paper has focused on what to do if there is not prior relevant information. This in no way stops pre-registering how the model of H1 will be constructed. One can pre-register for example "To model H1 for condition A, the SD of the half-normal will be half the effect for condition B." Pre-registering stops cherry picking the models of H1 one becomes fond of in the light of data. Bayes factors can be *B*-hacked just as *p*-values can be *p*-hacked (e.g. for both cases, how outliers are removed, whether a variable is in or out of the model, etc.), so pre-registering analytic protocols is just as valuable for Bayesians as frequentists.

The heuristics presented have partly been justified with the notion of severe testing: Although the heuristics sometimes use information from the very data used for testing a theory, they do so in a way that means strong evidence can still be produced against that theory. This claim seems to contradict Mayo (2018), who uses the notion of severe testing as an argument against Bayesian statistics (contrast Vanpaemel, in press). Mayo (2018) uses a notion of severe testing as a basis for understanding why cherry picking (i.e. selection effects in general) degrades evidence. She claims Bayesians struggle with explaining why selection effects degrade evidence. This is not so; in fact Bayesians are especially well placed

[7] For a uniform distribution, measure robustness by changing the upper limit of the uniform.

to explain when selection effects are bad and when they do not matter. Further, the Bayes factor also reveals why Popper's requirement of severity is related to evidence.

Popper (1963) defined a severe test as one where a predicted outcome was probable on the theory tested and improbable if the theory were false. Correspondingly, a Bayes factor is how much more probable the outcome is on the theory (or a model of it) versus H0 (for the examples we have considered). Thus, a test was severe if the Bayes factor departed considerably from 1. A Bayes factor measures strength of evidence defined as the amount by which one should change one's strength of belief. Thus, evidence goes hand in hand with severe testing. Take the outcome to be an obtained mean difference and its standard error. With researcher's degrees of freedom being used to cherry pick specific analytic decisions, there may be a similar probability of obtaining a given outcome on H0 as on H1; thus, the Bayes factor which took into account such selection effects as part of the data generating model would indicate there was little evidence (and that the test was not severe). Further, a Bayes factor indicates that selection effects caused by selecting what the precise model of the data is (what covariates are in the model etc.) in the light of what mean difference and standard error they produce, is different from selection in the form of optional stopping (see e.g. Rouder, 2014, Dienes, 2016, for discussion). The former degrades evidence and the latter does not.

This paper has discussed modelling of H1 and has not commented on the validity of the model of H0. Meehl (1967) argued that all point H0's are false (at least for correlational studies, but one could generalize his claim, cf. Greenland, 2017). So why would one want to test against a point H0? There is always a theoretically minimally interesting value, defining not a point null but a null interval (H0 specified say as a uniform, or a normal with small SD). This null interval can be hard to pin down exactly, but whenever the SE of the parameter is large compared to whatever the interval could be, the point null will be a good enough approximation to the interval. (And when the predicted scale of effect is in addition large compared to the SE, the Bayes factor will be informative.) So the point null is useful as it obviates the need to specify the null interval – and specifying the null interval, when done, should be done for objective reasons, which are often hard to come across. When a null interval can be approximately justified, it is easy to use in Bayes factors (e.g. Dienes, 2014, supplemental, section 6 for further discussion).

Greenland (2017) urged considering statistical models as thought experiments to guide intuitions and inference. Every assumption in a model of a psychological phenomenon will be an approximation, and we could have modelled the same phenomenon or theory in other ways. We can treat our models as conjectural, as things to be tested from any angle, with complete openness to revise in any direction, foreseen or not. We can test whether it is useful to have a parameter in the model by considering the scale of effects the parameter predicts or rules out. Without fixing that scale for some objective reason, there are no empirical grounds for removing a parameter. As Bayes factors take into account scale, they will often be relevant to testing models. This paper has provided some ways of thinking about what scale is relevant that may be helpful.

Box 4 Different philosophies for modelling H1

The way one approaches modelling H1 in a Bayes factor depends on one's philosophy of science.

1.Subjective Bayes factors (inspired by de Finetti, 1970). Probabilities are subjective and personal. Hence in representing a theory (e.g. the claim that a phenomenon exists) by a probability distribution for the different effect sizes predicted (the model of H1), one should consider the personal probabilities of a given individual (e.g. oneself, to make the outcome relevant to oneself). One can also carefully interview a range of experts, to obtain models that span from those skeptical of any but the smallest effects to those finding quite large effects plausible. The reader's predictions will hopefully roughly match up with one such model of H1. On this approach, one of the rock bottom processes of science is the rational persuasion of scientists until as a group they more or less agree about the support for a theory - even though each scientist has in effect their own personal model of the predictions made by a theory (models that should eventually converge).

2. Objective Bayes factors (inspired by Jeffreys, 1939). The claim that the precise predictions of a theory are a personal and individually varying matter will not fit everyone's philosophy of science. To escape having in principle a different model of H1 for every person, the most reassuring alternative may be having one model of H1 for almost all occasions - a default Bayes factor. Consider a two-group t-test. The within-group standard deviation defines an effect size regardless of original units yet an effect size that plausibly is the scale of effect that would obtain for many phenomena, to within a factor of 10. By having a default model of H1, post hoc cherry picking of one's model of H1 is avoided. Further, the stronger the evidence, the more robust the conclusion over different scale factors. That is, one need not fuss too much about the exact scale factor; just settle for a default. The problem is that scientists will always try to extract as many conclusions from data as they can, so they will reach down to the lowest degrees of evidence that inference will bear. Thus, inevitably we will deal with situations where the evidence is not overwhelming. In that case, default Bayes factors can be misleading, as we have seen in the paper.

3. Informed Bayes factors. Assume science is about testing theories by considering the objective relations between theory, assumptions and data (Popper, 1963). Each theory and set of assumptions is a conjecture (Popper, 1963); in that conjectural world, certain things follow, including the relative probability of different hypotheses, which we can assess by Bayes factors. The function of the model of H1 is to represent the predictions of a theory for reasons that are public and hence can be criticized. Thus, a theory should specify its predictions via well-justified and otherwise simple assumptions. Having constructed a draft model of H1, one still needs to make a plausibility judgment that the model adequately represents the predictions of the theory. On the one hand, in relying on a plausibility judgment, the informed Bayes factor is similar to a subjective Bayes factor. But that judgment should be treated not as an end in itself but as an indication of whether or not one can discover more constraints on predictions. For example, in using the room-to-move heuristic one might judge that the heuristic gives too large a scale factor. That judgment is an indication that, if you thought further, you may find objective reasons for why the effect should be smaller - and your job is to determine what those reasons are. On the other hand, the informed Bayes factor is similar to an objective Bayes factor in that scale factors are set for publicly available reasons. But in using an informed Bayes factor, unlike an objective Bayes factor, one must ensure that the model of H1 represents one's specific theory so that the measure of evidence given by the Bayes factor is relevant to one's theory. Thus, one cannot simply use default models of H1 without further thought about the relevance of the scale factor to the precise theory tested.

Acknowledgements.

Many thanks to Erin Buchanan Neil McLatchie, Mijke Rhemtulla, and Felix Schönbrodt for valuable comments.

References

Baguley, T. (2009). Standardized or simple effect size: What should be reported? *British Journal of Psychology, 100*, 603–617.

Balzarini, R. N., Dharma, C., Muise, A., & Kohut, T. (2019). Eroticism Versus Nurturance: How Eroticism and Nurturance Differs in Polyamorous and Monogamous Relationships. *Social Psychology, 50*, 185-200.

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., … Schönbrodt, F.D., ..., & Johnson, V. E. (2018). Redefine statistical significance. Nature Human Behaviour, 2, 6–10. doi:10.1038/s41562-017-0189-z

Caspar, E. A., Desantis, A. Dienes, Z., Cleeremans, A., & Haggard, P. (2016) The sense of agency as tracking control. *PLoS ONE 11*(10): e0163892. doi:10.1371/journal.pone.0163892

Cavanagh, K., Strauss, C., Cicconi, F., Griffiths, N., Wyper, A., & Jones, F. (2013). A randomised controlled trial of a brief online mindfulness-based intervention. *Behaviour research and therapy, 51*(9), 573-578.

Cortex (2019). https://www.elsevier.com/__data/promis_misc/PROMIS%20pub_idt_CORTEX%20Guidelines_RR_29_04_2013.pdf

De Finetti, B. (1970/translated 1975). *A theory of probability*. Wiley: Chichester.

Dickey, J. (1973). Scientific reporting and personal probabilities: Student's hypothesis. *Journal of the Royal Statistical Society. Series B (Methodological), 35*, 285-305.

Dienes, Z. (2008).*Understanding Psychology as a Science: An Introduction to Scientific and Statistical Inference*. Palgrave Macmillan. For associated Bayes factor calculator see: http://www.lifesci.sussex.ac.uk/home/Zoltan_Dienes/inference/Bayes.htm

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology, 5*, 781. doi: 10.3389/fpsyg.2014.00781

Dienes, Z. (2015). How Bayesian statistics are needed to determine whether mental states are unconscious. In M. Overgaard (Ed.), *Behavioural Methods in Consciousness Research*. Oxford: Oxford University Press, pp 199-220.

Dienes, Z. (2017a). https://www.youtube.com/watch?v=g9bIfZ4KqCQ See from 1 hr 20‘ to 1hr 27‘.

Dienes, Z. (2017b). https://www.youtube.com/watch?v=kWf65mMoJoU&t=682s

Dienes, Z., & Mclatchie, N. (2018). Four reasons to prefer Bayesian analyses over significance testing. *Psychonomic Bulletin & Review, 25*, 207-218.

Dienes, Z., & Seth, A. K. (2018). Conscious versus unconscious processes. In G. C. L. Davey (Ed.), *Psychology (BPS Textbooks in Psychology),* pp 262-323 . Wiley: Chichester

Etz, A., Haaf, J. M., Rouder, J. N., & Vandekerckhove, J. (in press). Bayesian inference and testing any hypothesis you can specify. *Advances in Methods and Practices in Psychological Science*.

Etz, A., & Vandekerckhove, J. (2018). Introduction to Bayesian inference for psychology. *Psychonomic Bulletin and Review, 25*, 5-34.

Gallistel, C. R. (2009). The importance of proving the null. *Psychological Review, 116*, 439–453.

Greenland, S. (2017). Invited Commentary: The Need for Cognitive Science in Methodology. *American Journal of Epidemiology, 186*, 639–645.

Gronau, Q. F., Ly, A., & Wagenmakers, E.-J. (in press). Informed Bayesian t-tests. *The American Statistician*.

Klaschinski, L., Schnabel, K., & Schröder-Abé, M. (2017). Benefits of power posing: effects on dominance and social sensitivity. *Comprehensive Results in Social Psychology, 2*, 55-67.

Jeffreys, H. (1939). *Theory of probability*. Oxford: Clarendon.

Lee, M. D., & Vanpaemel, W. (2018). Determining informative priors for cognitive models. *Psychonomic Bulletin & Review, 25*, 114 – 127.

Lindley, (1991). Comment on Aitkin "Posterior Bayes Factors." *Journal of the Royal Statistical Society. Series B (Methodological), 53*, 130-131.

Martin, J. R., & Dienes, Z. (in press). Bayes to the rescue: Does the type of hypnotic induction matter? *Psychology of Consciousness*,

Matzke, D., Ly, A., Selker, R., Weeda, W. D., Scheibehenne, B., Lee, M. D., & Wagenmakers, E.-J. (2017). Bayesian inference for correlations in the presence of measurement error and estimation uncertainty*. Collabra, 3:*25.

Mayo, D. G. (2018). *Statistical inference as severe testing: How to get beyond he statistics wars*. Cambridge: Cambridge University press.

McLatchie, N. (2018). http://www.neilmclatchie.com/bayes-robustness-regions/

Meehl, P. E. (1967). Theory testing in psychology and physics: a methodological paradox. *Philosophy of Science, 34*(2), 103-115.

Monin, J. K., Levy, B. R., & Kane, H. S. (2017). To Love is to Suffer: Older Adults' Daily Emotional Contagion to Perceived Spousal Suffering. *Journal of Gerontology Series B, 72* (3), 383–387.

Morey, R. D., Romeijn, J.-W., & Rouder, J. N. (2016). The philosophy of Bayes factors and the quantification of statistical evidence. *Journal of Mathematical Psychology, 72*, 6–18.

Nuijten, M. B., Wetzels, R., Matzke, D., Dolan, C. V., & Wagenmakers, E.-J. (2015). A default Bayesian hypothesis test for mediation. *Behavior Research Methods, 47*, 85-97.

Palfi, B., Parris, B. A., McLatchie, N., Kekecs, Z., & Dienes, Z. (2018). Can unconscious intentions be more effective than conscious intentions? Test of the role of metacognition in hypnotic response. *Cortex*, (Stage 1 Registered Report) https://osf.io/h2km3/

Parris, B. A., Dienes, Z., & Hodgson, T. L. (2013). Application of the ex-Gaussian function to the effect of the word blindness suggestion on Stroop task performance suggests no word blindness. *Frontiers in Psychology, 4*, 647.

Pearl, J., Glymour, M., & Jewell, N. P. (2016). *Causal Inference in Statistics – A Primer*. Wiley.

Popper, K. R. (1963). *Conjectures and Refutations: The growth of scientific knowledge*. London: Routledge.

Raz, A., Shapiro, T., Fan, J., & Posner, M. I. (2002). Hypnotic suggestion and the modulation of Stroop interference. *Archives of General Psychiatry, 59*(12), 1155–1161.

Rouder, J. N. (2014). Optional stopping: No problem for Bayesians. *Psychonomic Bulletin & Review, 21*, 301–308.

Rouder, J. N. (2019). On The Interpretation of Bayes Factors: A Reply to de Heide and Grunwald. https://doi.org/10.31234/osf.io/m6dhw

Rouder, J. N., Morey R. D., Verhagen J., Province J. M., & Wagenmakers E. - J. (2016). Is There a Free Lunch In Inference? *Topics in Cognitive Science, 8*, 520-547.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., and Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review, 16*, 225–237.

Schönbrodt, F. D.,Wagenmakers, E-J.,Zehetleitner, M.,Perugini, M. (2017). Sequential hypothesis testing with Bayes factors: Efficiently testing mean differences. *Psychological Methods, 22*, 322-339

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing Transparency Through a Multiverse Analysis. *Perspectives on Psychological Science, 11*, 702-712.

van Doorn, J., van den Bergh, D., Bohm, U., Dablander, F., Derks, K., Draws, T., … Wagenmakers, E. (2019, January 23). The JASP Guidelines for Conducting and Reporting a Bayesian Analysis. https://doi.org/10.31234/osf.io/yqxfr

Vanpaemel, W. (in press). Strong theory testing using the prior predictive and the data prior. *Psych Review*,

Woody, E. (2011). An SEM Perspective on Evaluating Mediation: What Every Clinical Researcher Needs to Know. *Journal of Experimental Psychopathology, 2*, 210–251

Salvador, A., Berkovitchd, L., Vinckiera, F., Cohen, L., Naccache, L., Dehaene, S., & Gaillard, R. (2018). Unconscious memory suppression. *Cognition, 180*, 191–199.

Yzerbyt, V., Muller, D., Batailler, C., & Judd, C. M. (2018). New recommendations for testing indirect effects in mediational models: The need to report and test component paths. *Journal of Personality and Social Psychology, 115*, 929–943.

Ziori, E., & Dienes, Z. (2015). Facial beauty affects implicit and explicit learning of men and women differently. *Frontiers in Psycholology, 6*, 1124. doi: 10.3389/fpsyg.2015.01124